International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 5, Issue 2, Apr 2015, 41-52

OPINION MINING OF REVIEWS DISTRIBUTED OVER E-COMMERCE SITES

REKHA MORE & EKTA UKEY

Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Maharashtra, India

ABSTRACT

An enormous growth in web technology allow customers to post reviews about products they bought. These reviews helps customers as well as manufacturer. But there are thousaands of reviews of customers related to one product. Moreover these reviews are in a unstructured or semistructured form. So it is difficult for the customers as well as manufacturer to have an idea about the product from these large reviews. In this paper system analyse the existing opinion mining technique and then see the implementation of enhanced feature based opinion mining technique which analyse the customer reviews distributed over different e-commerce sites. For summerising and analysing reviews two statistical techniques are used namely bayesian probability and frequency distribution. Frequency distribution represent the result in graphical form, sofrequency based results can be easily understood by all customers. And bayesian probability verify the results of frequency.

KEYWORDS: Opinion Mining, Review Analysis, Frequency Distribution

INTRODUCTION

Automatic detection and analysis of customer reviews on the web are very important for market research and customer relationship management. Many people always asks their friends or relatives while buying various household appliances and various other daily used items such as kitchen cutlery, mobile phone, other electronic gadgets etc. So they take a rational decision about different buying patterns. The opinions, sentiments, recommendations and consultancy provided by the concerning persons during decision making has become the topic of great worth more over these opinions may be multiple and opposite which must be integrated and summarized. Different techniques lead to the analysis of web documents and focuses on customer reviews available on different platforms such as discussion forum, Blogs and companies sites and other social media, reflecting positive or negative experiences and opinions of customer about different brands of product and services[1]. Customers can get feedback from a company; the position can take criticisms and suggestions which include hundreds and thousands of comments that need to be treated by some intelligent system for automatic summarization and classification. There are various methods for opinion mining; some well-known methods in respect with opinion mining are association rule mining, machine learning, NLP and text mining.

In this research main focus on feature based opinion mining using statistical techniques to find the opinion of the customer reviews and analyse it word by word, then classifying each word as positive or, negative or neutral, which is nothing but polarity of opinion text(positive/negative/neutral). After determining the polarity, calculate the frequency distribution and bayesian probability of polarity for the entire product.

Finally, summerise and analyse customer reviews distributed over different e-commerce sites. Frequency distribution and bayesian statistics previously used for mining customer reviews only for particular e-commerce site. But in our research we use this technique to perform opinion mining of reviews distributed over different e-commerce sites. The

basic goal of our research is to represent the results in simple form which are easy to understand for customers and manufacturer.

There are many opinion techniques are available which helps to analyse the customer reviews. But there are certain limitation while using these techniques. Like Association Rule mining [2] techniques proposed by Hu and Liu consider only frequent features addressed by customer, and ignored infrequent feature. Qui et al. proposed a novel mutual reinforcement approach[3] to deal with the feature-level opinion mining problem. this approach show the relationship between opinion and feature, but relationship between opinion and feature are so complex that the error will increase with each iterations. Popescu and etizoini also investigate the same problem, Thealgrithm [3] only reckon noun/noun phrase as the candidate features. It determines whether a noun/noun phrase is a feature by computing the Point-wise Mutual information [4] (PMI) score between the phrase and class discriminators, it calculate it calculates the PMI by searching the web. Querying the web is time-consuming. In 2009 Qui et al proposed double propogation technique which state that opinion words can be recognized by identified features, and features can be identified by known opinion words. So the extracted opinion words and features are utilized to identify new opinion words and new features, which are used again to extract more opinion words and features. Zhang and Liu improved the double propogation [5]. The approach used two patterns which is based on part-whole patterns and "no" patterns to increase the recall and precision.

The major tasks of feature based opinion mining are - (1) to identify the products features in review, (2) to determine opinion expressed by the reviewer (positive, negative or neutral), (3) summarize discovered information. Some of the mining system follow ontology based feature based opinion mining, domain-specific opinion extractin, and some other automated techniques for feature based opinion mining [6]. Almost all of these follow some basic steps to achieve their goal. These steps are shown in following Figure.

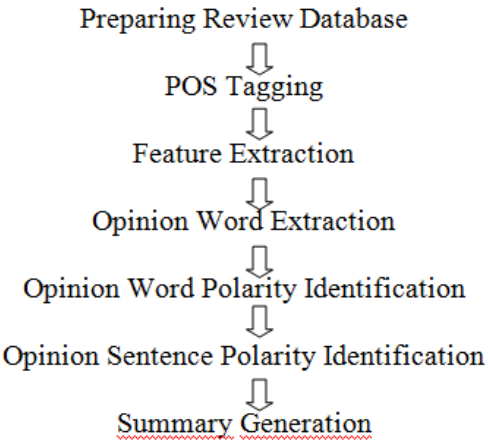


Figure 1: Basic Steps for Feature Based Opinion Mining and Summarization

METHODS

The idea of Enhanced feature based opinion mining Architecture was first proposed by Naveed Anwar, Ayesha Rashid in "Naveed Anwar, Aayesha Rasheed, Sayeed Hasan,"Feature Based Opinion Mining of online free format customer reviews using frequency distribution and Bayesian statics", pages 378-385,2013, IEEE as shown in figure 2. But according to the authors "The proposed architecture can achieve the feature based opinion mining in theory, but theyhaven't tested its advantages and disadvantages in practical application. To design more general and more

comprehensive feature based opinion mining model, need of more efforts from the circles of academia and industry are required.".

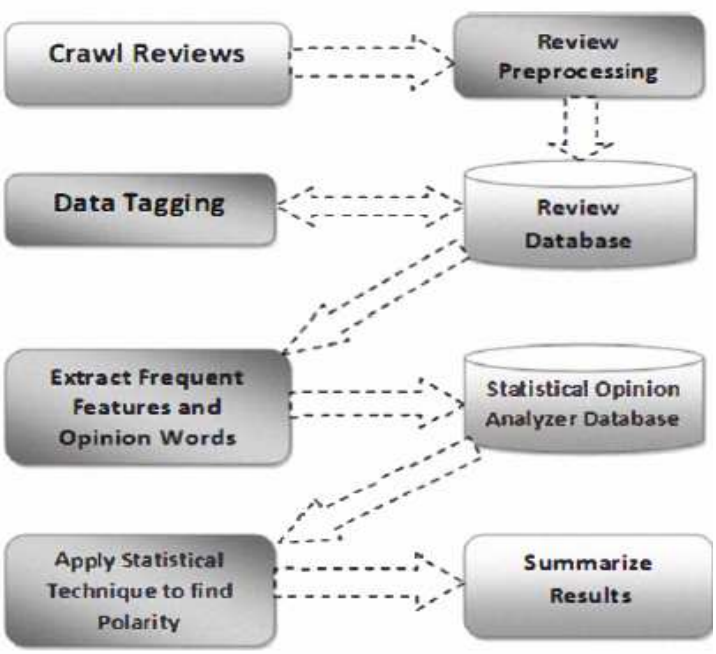


Figure 2: Statistical Opinion Analyzer Flowchart

We have successfully implemented the above system. Figure 3 shows the flow of project, where system first, extract the customer feedbacks from the websites and add them to the review database. Second, use Part of Speech (POS) Tagging from natural language processing to tag these reviews; third extract all frequent features and associated polarity (positive, negative) for all the products and stored in database. Finally, for each product, the system finds the polarity feature wise and overall polarity by applying Bayesian probability and Frequency distribution. Then find the polarity of each product with a total of polarity with the frequency distribution by calculating the positive and negative comments. Also system compare the mining results of e-commerce sites for better analysis.

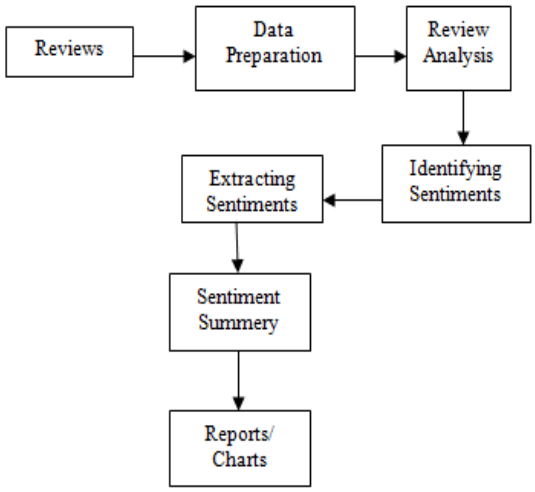


Figure 3: Flow of Enhanced Feature Based Opinion Mining

Flow of the Project

Step 1: Formation and Data Collection

Customers may comments on three types of format, free format, pros and cons format and mix format. We used free type format reviews in our research that the reviewer can write freely, i.e., no separation of Pros and Cons. The main idea is that every adjective is used to modify a feature, no matter what it is the whole entity or a feature of the entity. The method not only considers the noun /noun phrase but also the verb/verb phrase as the features.

Step 2: Review

All the irrelevant information removed from reviews file to make it easy for further use. Data pre-processing include cleaning of data by removing numbers, HTML tags, website's own information given on the web pages, symbols, spelling mistakes and the extra useless information e.g. date, name of the reviewer or reference of any third person who is not seems to be relevant. Part of speech tagging (POS) from NLP can be defined as assigned text to each word in review on the basis of characteristics of the word and the context in which it occurs.

Step 3: Pre- Processing

In this step a transaction file is created for the generation of frequent features along with potential opinion phrases and finds the polarity of each opinion in the context of its associated feature in a particular review sentence. Only the features which are frequently commented by customers are extracted.

Step 4: Finding Opinion Phrases and Polarity

Prediction of the orientation of opinion sentence is to identify complete sentence as positive or negative. And the prediction of the orientation of the overall product is to identify overall sentiments expressed by authors against single product. The system find the polarity ofany product along with its feature so to find the polarity of overallproduct, system find the frequency distribution by calculating the positive and negativenopinion then results areshown in graphical form.

Calculating Frequency Distribution

Number of occurrences in the given domain is called the frequency of that domain. Frequency of positive as well as negative opinions ofnearby feature is calculatedPositiveopinion represent that how many times customer commented on the product positively and same for negative opinion.

Calculating Bayesian Probability Distribution

After calculating frequency distribution calculating bayesian probability of product, deliver more accurate and better result.

Bayesian probability is the most accurate and updated probability in statistics, which are used to provide accurate results and true predictions. The formula of Bayesian statistics is

Let E be an event in a sample space S, and let (A1, A2....., An) be disjoint events whose union is S. then for K = 1, 2, 3 , n.

$$P(A_K/E) = \frac{P(A_K)P(E|A_K)}{P(E)} \quad \text{.....(1)}$$

The above equation is called Bayes' rule or Bayes' formula.

If we think of the events even A1, A2..., An as possible causes of the event E then Bayesian formula enables us to determine the probability that a particular one of the A's occurred given that E occurred. Each term in Bayes' theorem has a conventional name. A is a hypothesis, and D is the data:

- P (A) is the prior probability or marginal probability of A. it is "prior" in the sense that it does not take into account any information about B.
- P (A/E) is the conditional probability of A given E. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- P (E/A) is the conditional probability of E given A. In the denominator of the formula P (E) let E be an event in a sample space, and let A1, A2, A3 be mutually disjoint events whose union in S then

P (E) =P (Al). P (E/Al) P (E/A2) + ... +P (An). P (E/An) …… ……....(2)

This equation is known as the law of total probability. Where we emphasise that the sets (A1, A2, A3 .... An) are pair wise disjoint and their union is all of S.

Step 5: Analytics and Report

Analytics and report gives overview of the current opinion mining project. It may represent graphs, statistical reports and simple reports. Analytics refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.

RESULTS

Scenario which Show Positive Keywords

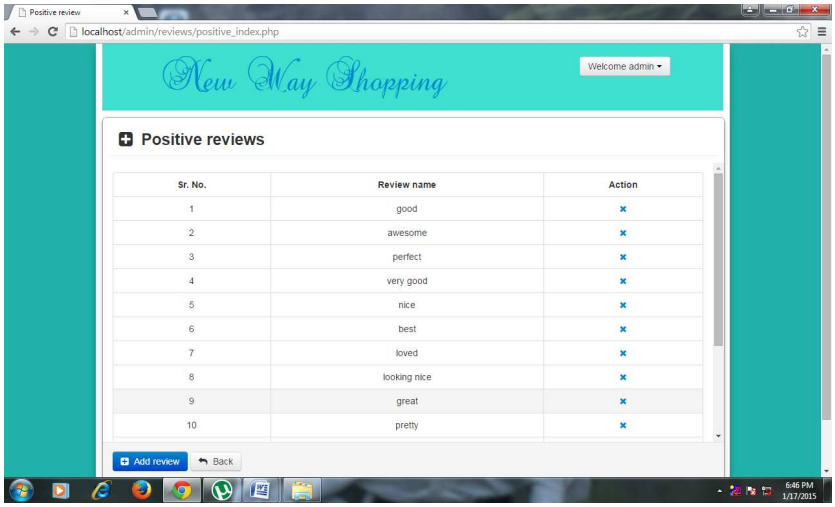


Figure 4

Scenario Which Show Negative Keywords

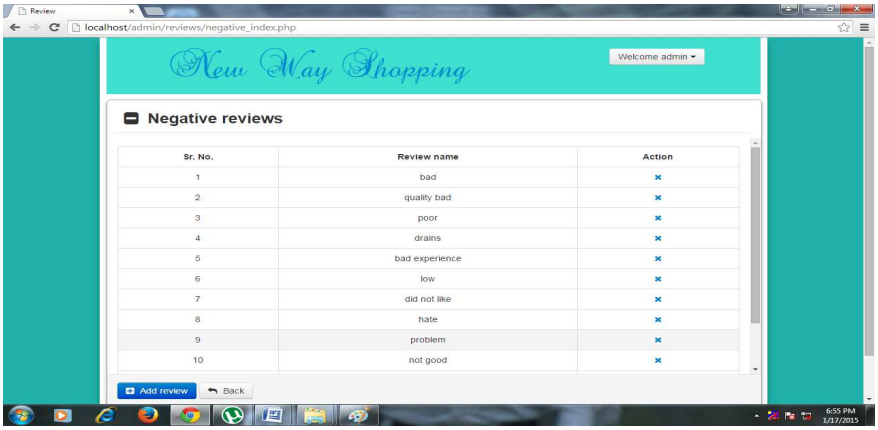


Figure 5

Scenario which Show Neutral Keywords.

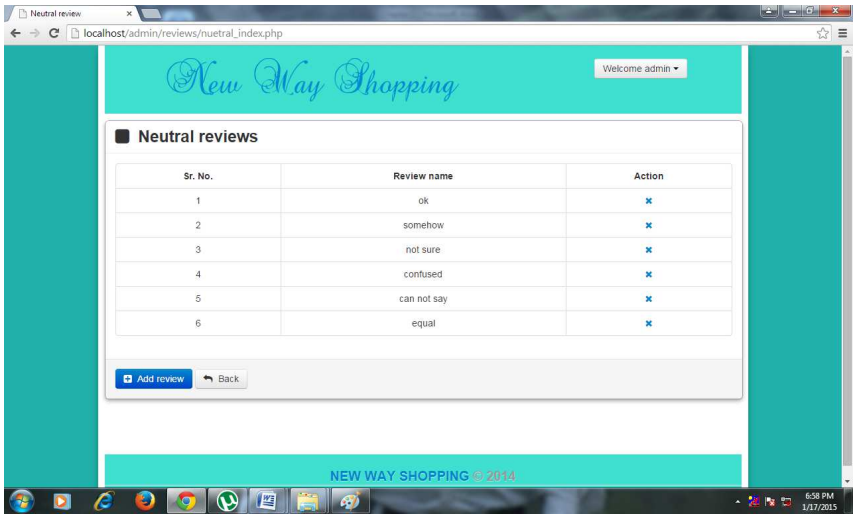


Figure 6

Scenario for Opinion Mining Result

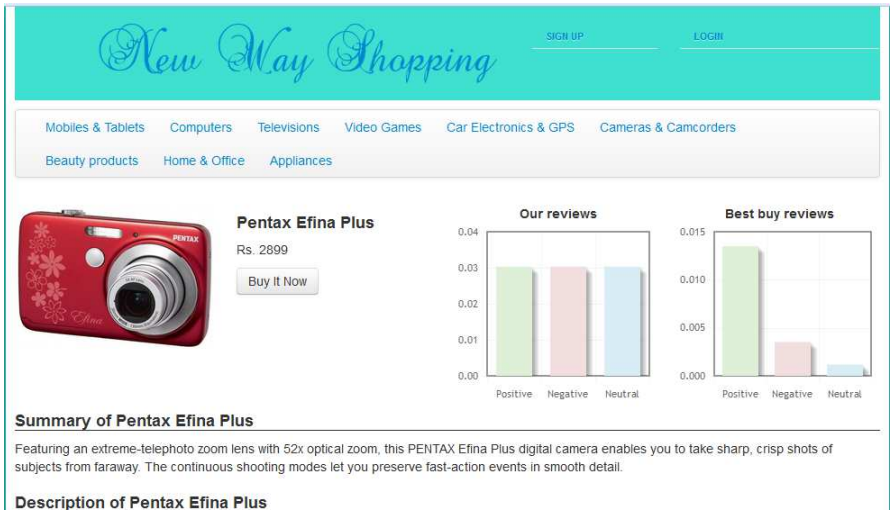


Figure 7

Scenario for Frequency Distribution Graph Which Show Opinion Mining Result Using Bayesian Probability

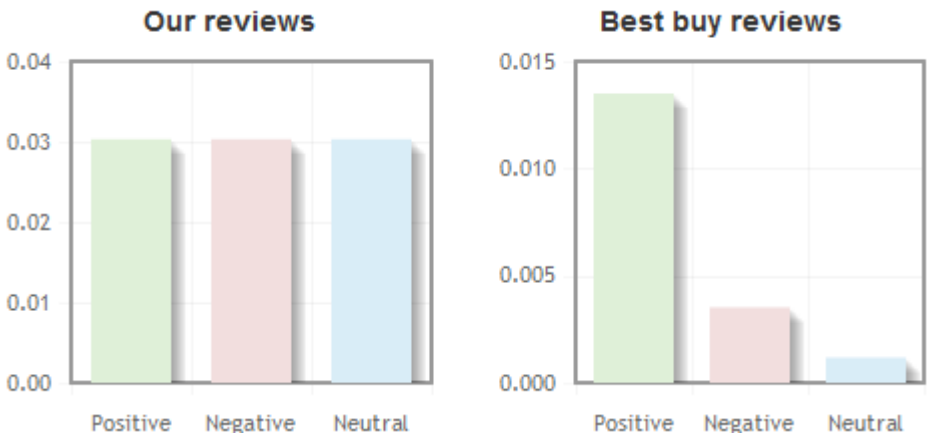


Figure 8: Scenario for Frequency Distribution Graph

Scenario for Comparison Result of Two Products

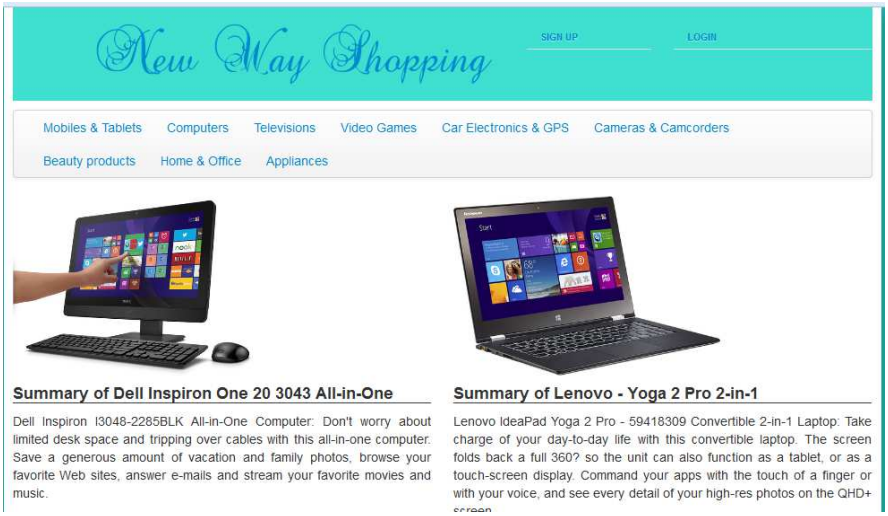


Figure 9: Scenario for Comparison Result of Two Products

Example

A person want to purchase tablet from new way shopping. but before purchasing tablet he checked the reviews related to different companies tablet and then he will decide which one to purchase. A feature based system help customer to take decision while purchasing tablet. Following are the comments related to Asus fonepad which consists of our reviews with frequency distribution graph:

- Good picture quality.....
- Good sound quality.....
- The audio and call quality is terrible when using the phone directly.
- I like the the phone / tablet combination! I can talk and look up data at the same time on one device! I say well done to Asus.

- I bought asus fonepad 7 in august2014 but now its google play store is not working.
- Excellent tab.
- I can't take selfie in this fonepad using beauty plus apps it always saying that turn on the fill flash mode what can i do?
- For phone ring not audible.
- Ok.

The following table consists of positive, negative and neutral words are as follows

Table 1

Positive	Negative	Neutral
Good	Terrible	Ok
Like	Not working	-----
Excellent	Can't	-----
----	Not audible	-----

Solution

Total word counts in above reviews are 93.

Extract frequent features and opinion words →Count frequency

$$P(A/B) = \frac{P(A).P(B/A)}{P(B)}$$

i.e $$posterior = \frac{prior * likelihood}{evidence}$$

In some cases evidence is ignored.

So, Probablity of "good" in total reviews is calculated by following formula

$$P(good/positive) = \frac{P(good).P(positive/good)}{p(positive)}$$

= (2/93)*(0.333) = 0.007161

Probablity of "like" in total reviews is calculated by following formula

$$P(like/positive) = \frac{P(like).P(positive/like)}{p(positive)}$$

=(1/93)*(0.333) = 0.00358

Probablity of "excellent" in total reviews is calculated by following formula

$$P(excellent/positive) = \frac{P(excellent).P(positive/excellent)}{p(positive)}$$

=(1/93)*(0.333) = 0.00358

Next, Probability of "terrible" in total reviews is calculated by following formula

$$P(\text{terrible}/\text{negative}) = \frac{P(\text{terrible}).P(\text{negative}/\text{terrible})}{p(\text{negative})}$$

= (1/93)*(0.333) = 0.00358

Next, Probability of "not working" in total reviews is calculated by following formula

$$P(\text{not working}/\text{negative}) = \frac{P(\text{not working}).P(\text{negative}/\text{not working})}{p(\text{negative})}$$

= (1/93)*(0.333) = 0.00358

Probability of "not audible" in total reviews is calculated by following formula

$$P(\text{not audible}/\text{negative}) = \frac{P(\text{not audible}).P(\text{negative}/\text{not audible})}{p(\text{negative})}$$

= (1/93)*(0.333) = 0.00358

Probability of "cant" in total reviews is calculated by following formula

$$P(\text{cant}/\text{negative}) = \frac{P(\text{cant}).P(\text{negative}/\text{cant})}{p(\text{negative})}$$

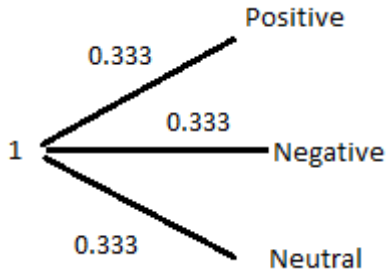
= (1/93)*(0.333) = 0.00358

Probability of "ok" in total reviews is calculated by following formula

$$P(\text{ok}/\text{neutral}) = \frac{P(\text{ok}).P(\text{neutral}/\text{ok})}{p(\text{neutral})}$$

= (1/93)*(0.333) = 0.00358

Bayesian network for positive, negative and neutral are shown below



P(positive) = P(good/positive)+P(excellent/positive)+P(like/positive)

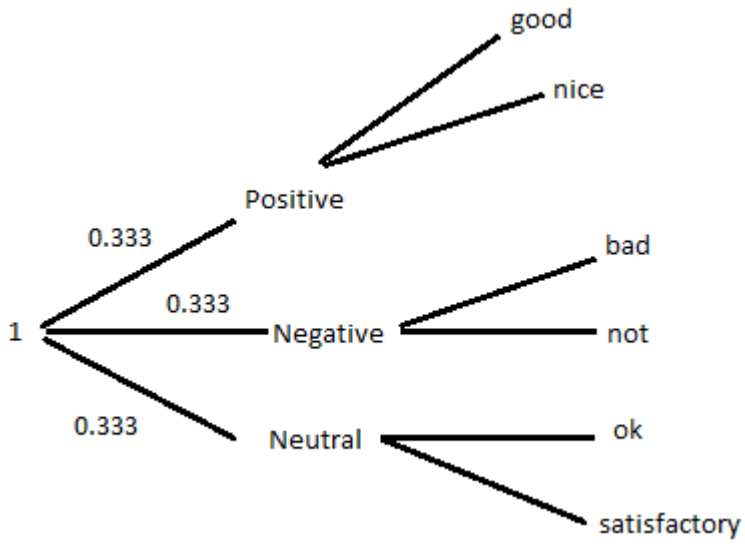
= 0.007161 + 0.00358 +0.00358 = 0.014321

P(negative) = P(terrible/negative)+P(not working/negative)+P(not audible)

= 0.00358+0.00358+0.00358 = 0.01074

P(neutral) = P(ok/neutral) = 0.00358

So the Bayesian networks for all keywords are as follows:



CONCLUSIONS

Feature based mining system is a supervised information extraction system which extracts fine-grained features. Associates opinion properly from online product reviews to identify product features with fairly good accuracy. In this thesis main aim is mining customer opinions. For analysing opinion Bayesian probability is used which gives most accurate result in statistical mathematics. And frequency distribution shows graphical interface which analyze the customer opinions of particular product by viewing the negative and positive frequency of opinion along with its probability about product and customer can easily judge the product as well as online companies. Online companies can perform automatic analysis on multidimensional opinion. Feature based system makes complete analysis on customer review document and present the comparative analysis of opinion between different sites or Blogs in an automatic and efficient manner.

ACKNOWLEDGEMENTS

Authors are greatful and thankful for the Bestbuy web services for letting the acsess to the product data.

REFERENCES

1. Naveed Anwar, Aayesha Rasheed, Sayeed Hasan,"Feature Based Opinion Mining of online free format customer reviews using frequency distribution and Bayesian statics", pages 378-385,2013.
2. Kavitha Murugeshan, Neeraj RK "Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.
3. Jian Ma, Wei Xu ; Yong-hong Sun ; Turban, E "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on (Volume:42, Issue: 3), 2012.
4. Ruowu Zhong, Huiping Wang,"Research of Commonly Used Association Rules Mining Algorithm in Data

Mining" Internet Computing & Information Services (ICICIS), 2011 International Conference , ISBN- 978-1-4577-1561-7, Page(s): 219 – 222,2011.

5. Jingjing Kang, Xiaoyong Du, Tao Liu, He Hu," Automatic Domain Terminology Extraction Using Graph Mutual Reinforcement", 11th International Conference, WAIM 2010, Jiuzhaigou, China, July 15-17, 2010. Proceedings, ISBN- 978-3-642-14245-1, Volume-6184, pp 656-667,2010.
6. Lei Zhang, Bing Liu, Suk Hwan Lim, Eamonn O'Brien-Strain," Extracting and Ranking Product Features in Opinion Documents", Coling 2010: Poster Volume, pages 1462–1470, Beijing, August 2010.